A Platform for IoT and Social Big Data

Takahiro HARA
(hara@ist.osaka-u.ac.jp)
Osaka University
Main theme

People as Sensors

Keywords

IoT (Internet of Things)  Social Media (SNS)  Big Data

A Platform for IoT and Social Big Data
Background

◆ Trend: Big data applications with social benefits
  ➢ Ex) disaster management, crime prevention, urban design, safety/autonomous driving

◆ IoT and Social media play important role as data sources.

NHK Special: “Disaster Big Data”
IoT and Social media

IoT (M2M, Sensors, etc.) and Social media (Twitter, Facebook, flickr, etc) will not lead us to Cold (Human-less) World.

Rather, make people more visible to ICT systems, i.e., better involved in Big Data applications.

- IoT: Mobility/Behavior tracking, Situation (context) awareness, etc.
- Social media: Event detection, Sentiment analysis, Trend analysis, etc.
Flow of IoT/social big data analysis

Science Data  |  Enterprise Data  |  Web/Social Media Data  |  Sensor Data  |

Data Collection

Data Analysis

Result (Data/Model)

Pattern mining
Real-time prediction etc.
Characteristic of IoT/Social big data

IoT/Social media enable real-time real-world analysis

- **Characteristic:** Data come from different data sources. Ex) sensors, service logs, social media data

Two main challenges

- **Parallel computing framework for IoT/social Big Data**
  - IoT data: have various spatial/temporal properties.
  - Social data: require heavy computation.

- **Open-data framework for sharing IoT/social Big Data**
Frameworks for sharing IoT Big Data

Existing frameworks

- **SODA (Keio Univ.):** Framework for leveraging social big data for open smart cities
  - Tools for crowdsensing, time-series analysis, indoor positioning, etc.
- **TAREEG (Univ. Minnesota):** Web-service for sharing a large volume and variety of spatial data
  - MapReduce-based techniques for efficient data processing

What are missing?

- No framework aims to share social media analytical results (**Social Sensor Data**), while useful for other applications.
Social media as data sources

- Social media (e.g., twitter) are important data sources for detecting real-time events.

- **Current situation for disaster management:**
  Over 20M messages posted during Hurricane Sandy and Haiti earthquake, only 100K messages were actually processed by the government agencies.
Motivation

◆ A large number of studies on Social media mining
  ➢ Event detection, Trend analysis, Sentiment analysis, etc.
  **Social Sensor!**
  ➢ Most studies do not reuse the analytical results. (sole app)

◆ **Goal:** Building a platform integrating (ordinary) sensor data and social sensor data
  ➢ Reuse of social media analysis

Ex. Event detection from Twitter posts

Useful for developing Big Data applications using people as sensors.
Research challenges

◆ Creating social sensor data from Social media data

**Sensor reading**

**Social Media Analysis**
- Local event detection,
- Trend/Sentiment analysis

**Location**

**Geocoding**
- Attaching location info. to Social media data

**Time**

◆ Developing a framework for sharing (reusing) Social media analytical result (**Social Sensor data**)
Creating social sensor data from Social media data
Creating Social Sensor data

Step 1: Generating Social Sensor reading

(Ex) Local event detection using geotagged tweets

Using spatiotemporal locality of geotagged tweets.

1. Find clusters where many users posted tweets in a certain period of time.
2. Detect key term co-occurrences in each cluster that represent the event.

Geotagged tweets
Creating Social Sensor data

Step 2: Geocoding (location identification)

Learn patterns when a place name in a tweet indicates user’s location. Massive geotagged tweets can be used as training data.

Training phase

Geocoding

Geotag

Location by geocoding

Learn features when locations by geocoding and geotag are close (positive) and far (negative).

Classification phase

Geocoding

Location by geocoding

Add geotag when SVM classifies tweets as positive.
Reusing (sharing) results of Social media data analysis
How to reuse social media analysis?

**Basic: Reuse of analytical result**

- **Challenge:** Making result accessible (e.g., by search)
  - Formal (type) definition of the result as SS data
  - Ex) SS name, data source, analytical methods.

**Advanced: (Partial) Reuse of analytical method**

- **Challenge:** Making procedures accessible and editable
  - Code sharing in an open source manner
  - Platform independent language

**Code sharing (Open source)**

Social media data

Analytical procedure (program)

```
class createData()
{
    public void getSNS()
    {
        function for getting SNS posts
    }
    ....
    main()
    {
        ....
    }
```

Analytical result (SS data)

- SS Name: Local Event
- Source: Twitter
- Method: ...

- L1, T1, Fire
- L2, T2, Surprise Live
- L3, T3, Car accident
- L4, T4, Guerrilla rain
Our idea: Dividing code into 3 files

- **Social Sensor Type Def. (SSTD):** XML
 _defs of function names for SS Data generation
  ```
  public void getSNS()
  analyzeSNS()
  ```

- **SS Function Def. (SSFD):** Java
  Procedures of the functions
  ```
  public void getSNS(){
    SNS post getting procedure
  }
  public object analyzeSNS(){
    topic generation procedure
  }
  ```

- **SS Output Config. (SSOC):** XML
  Output DB names and others
  ```
  Database config.
  etc.
  ```

---

**Runtime Program for generating SS Data**

```java
class createData(){
    public void getSNS(){
      function for getting SNS posts
    }

    public object analyzeSNS(){
      topic detection procedure
    }

    public static void main(){
      ...
    }
}
```
Our idea: Platform to generate SS Data

Based on the uploaded program, the Job Scheduler generates Social Sensor Data (SSD).

SS Data Generator

"Runtime program"

Syntactic Parsing

SSOC

SSFD

SSTD

Web interface

Job Scheduler

Registration of info on program execution

Trigger

Request

SS Data

Social Sensor (SS) DB

SNS posts data repositories

Web interface

SS DB Creation Acceptor

SSOC

SSFD

SSTD

17 JOFOE 2016
Summary: What we did so far

◆ Developed the minimum system functions to share Social Sensor data (Social media analysis).
  ➢ i.e. program upload, execution, storing the result, making program and result available.

◆ Of course, many things to be considered...
Future directions (Open issues)

◆ How to collect missing data necessary for analysis?
  ➢ Solution: Crowdsensing: Giving tasks for data collection to ordinary people.
  ➢ Benefit: Useful for other applications

◆ Ownership/Copyright problem
  Who is the owner of SS data generated from multiple data sources, e.g., tweets? (message authors? analyst?)
  ➢ Solution: Development of lows + Data traceability
  ➢ Benefit: Encouraging open data and data distribution

◆ Credibility problem
  How can we know/guarantee the credibility of SS data?
  ➢ Solution: Standardization of performance metrics + automatic generation
  ➢ Benefit: Making SS data comparable + Quality improvement

◆ (Privacy problem)